# Predicting The Effects of Chemical–Protein Interactions On Proteins Using Tensor Factorisation

**Sameh K. Mohamed**[1,2], **Aayah Nounu**[3]
[1]**Data Science Institute, NUI Galway, Galway, Ireland**
[2]**Insight Centre for Data Analytics, NUI Galway, Galway, Ireland**
[3]**MRC Integerative Unit, Bristol Medical School, University of Bristol, Bristol, UK**

**Abstract**

Understanding the different effects of chemical substances on human proteins is fundamental for designing new drugs. It is also important for elucidating the different mechanisms of action of drugs that can cause side-effects. In this context, computational methods for predicting chemical–protein interactions can provide valuable insights on the relation between therapeutic chemical substances and proteins. Their predictions therefore can help in multiple tasks such as drug repurposing, identifying new drug side-effects, etc. Despite their useful predictions, these methods are unable to predict the different implications —such as change in protein expression, abundance, etc,— of chemical–protein interactions. Therefore, In this work, we study the modelling of chemical–protein interactions' effects on proteins activity using computational approaches. We hereby propose using 3D tensors to model chemicals, their target proteins and the effects associated to their interactions. We then use multi-part embedding tensor factorisation to predict the different effects of chemicals on human proteins. We assess the predictive accuracy of our proposed method using a benchmark dataset that we built. We then show by computational experimental evaluation that our approach outperforms other tensor factorisation methods in the task of predicting effects of chemicals on human proteins.

## Introduction

Understanding the different effects of chemical substances on human proteins is fundamental for designing new drugs. It is also important for elucidating the different mechanisms-of-action of current drugs that can cause unwanted side-effects [(1)]. This encouraged researchers to investigate the different chemical–protein interactions and their effects on the protein activity in living systems. Chemicals can have different types of effects on their target proteins such as change of expression, abundance, secretion, etc. These different effects then play various roles in the mechanism-of-action of chemicals in living systems. Therefore, understanding the chemical–protein interactions with their respective effects is crucial to elucidating the mechanism-of-action of therapeutic chemical substances.

The process of investigating chemical–protein interactions and their effects commonly involves 'omics approaches such as mass spectrometry which is used for the generation of proteomic data. This method is mainly used to identify off-target or non-canonical targets of chemicals (drugs) that may be unknown [(2, 3)]. Other lab approaches include phenotypic screening such as work carried out by Iljin et. al. [(4)] whereby 4910 drug-like small molecule compounds were tested against prostate cancer cell lines to identify which ones affected cell proliferation [(4)]. Despite the insightful findings of such approaches, they are laborious, time-consuming and resource-consuming processes.

This encouraged the development of different computational approaches to inform and assist laboratory experimentation of chemical–protein interactions [(5–7)]. These approaches enabled predicting the most plausible chemical–protein interactions with high accuracy and efficiency [(6, 7)]. However, all the current computational approaches are only focused on predicting the existence of chemical–protein interactions, and they do not provide any information on the effects of these interactions.

In this study, we extend the design of traditional chemical–protein interaction computational approach to allow them to encode the different types of effects caused by these interaction on target proteins. We model the information on chemicals, their target proteins and associated interaction effects as 3D where we can easily apply tensor factorisation methods to infer new unknown chemical–protein interactions' effects.

Tensor factorisation methods have been widely adopted in different biological tasks including drug–target interaction prediction [(7, 8)], drug side-effect prediction [(9, 10)], protein biological functions prediction [(11)], etc. Tensor factorisation approaches were then used to generate vector representations of biological entities to provide predictions of their unknown associations to other entities.

In the context of our study, we use tensor factorisation methods to generate embeddings of chemicals, their targeted proteins and corresponding chemical effects. We then use these embeddings to predict new possible chemical–protein interactions and their associated chemical effects. We use a multi-part tensor factorisation approach that models tensor objects' embeddings using multiple tensors and we show that this approach has better accuracy than other tensor factorisation approaches. To the best of our knowledge, this work is the first computational method that considers the context of chemical effects in predicting chemical–protein interactions. Therefore, we only compare our method to other tensor factorisation approaches.

We build a benchmarking dataset that consists of known chemical–protein interactions with their associated chemical effects extracted from the Comparative Toxicogenomics Database (CTD) [12, 13]. We then show by computational experimental evaluation on this benchmark that our proposed multi-part tensor factorisation approach outperforms all other approaches in the task of predicting chemical–protein interactions and their associated chemical effects.

The rest of this study is outlined as follows: the background section presents a brief background on the studied problem and proposed approach. The materials section discusses the benchmarking dataset used in this study. The methods section discusses the design details of our approach. The results section presents the experimental setup of our evaluation benchmark and the outcome results. The discussion presents the findings of this study and discuss the challenges and limitations of the proposed approach.

## Background

In this section, we discuss the preliminary concepts and notations that we use through this study. We first discuss the problem of the different systematic approaches used for drug repurposing. We then discuss the tensor factorisation model and other notations that we use during the training and evaluation of our approach.

***Drug repurposing approaches.*** Drug repurposing is the use and effectiveness of well-known drugs for alternative diseases other than the disease it was originally designed for [14]. This is a more cost-effective and time-efficient process than developing new drugs, as it bypasses the need for the drug to go through Phase I of the clinical trials since it already has a known safety profile [15]. For this reason, more systematic approaches have been developed including both computational and experimental approaches.

The first example of a computational approach includes signature matching-comparing transcriptomics, chemical structures or adverse drug effects between two different drugs. Increased similarity in these signatures indicates similar targets [16–18]. Other examples include the use of genome-wide association studies (GWAS) whereby an association of a genetic loci with one disease may also be associated with other diseases. This shared association may indicate the potential to use the same drug to treat the diseases [19]. Another approach involves retrospective analysis of electronic health records which has been useful in alluding to the repurposing of drugs. This approach has been used for repurposing drugs like *Sildenafil Citrate* [20]. Furthermore, it was also recently used for repurposing *Aspirin* which was originally used for cardiovascular diseases but has now been recommended by the US Preventive Services Task Force for the chemoprevention of colorectal cancer [21].

In this work however, the scope of our study is focused on computational approaches that utilize information about the drug–protein interactions and their chemical effects on the protein activity.

***Tensor factorisation.*** Scalars are singular numerical values, vectors are one dimensional numerical arrays, matrices are two dimensional numerical arrays, and tensors are numerical arrays with three or more dimensions. In our study, we focus on tensors with only three dimensions. In this case, a tensor cell represents the interaction between three components from the different tensor dimensions. This interaction commonly denotes the likelihood of a joint association between the three components which are represented by the cell. In practice, each dimension of a tensor data model represents components from a specific type such as chemicals or proteins in our case. For example, let us assume that we have a 3D tensor of chemicals, effects and protein targets. The value of the cell corresponding to the three components (*1-nitropyrene, decreases_expression, LRRC17*) will be 1 if the chemical "*1-nitropyrene*" has the effect "*decreases_expression*" on the protein target "*LRRC17*" and 0 otherwise.

The objective of the procedure of tensor decomposition *i.e.* tensor factorisation, is to complete all cell values in

an incomplete tensor using a set of initial known cell values. This procedure is achieved by learning numerical representations of different tensor objects. The representations (*i.e.* embeddings) are then used to provide scores for any given tensor combination.

Let $\mathbf{M}$ be a three dimensional tensor, where the three dimensions represent objects of different sets $X$, $Y$, $Z$. Any element $(i, j, k)$ in the tensor represents the interaction between the components $i \in X$, $j \in Y$, and $k \in Z$. We denote the weight of this interaction using $\eta^{\mathbf{M}}(i, j, k)$. In this study, we use a tensor $M$ with elements of the three sets: chemicals ($\mathbf{C}$), effects ($\mathbf{E}$), and target proteins ($\mathbf{P}$). The objective of tensor decomposition then is to complete the tensor values such that the weight of any interaction of a true known chemical effect on a protein target is larger than all other known false combinations. This can be defined as follows:

$$\forall(c, e, p)\forall(c, e, p)' \quad \eta^{\mathbf{M}}(c, e, p) > \eta^{\mathbf{M}}(c, e, p)' \tag{1}$$

where $c \in \mathbf{C}, e \in \mathbf{E}, p \in \mathbf{P}$, $(c, e, p)$ is any known true combination of a chemical, an effect and a target protein such that the chemical $c$ has the effect $e$ on the protein $p$, and combinations $(c, e, p)'$ represent any other false combinations. This objective is achieved using a multi-phase procedure where the model iteratively learns the missing scores by processing each of the initially known tensor combinations. In this work, we use the learning procedure of knowledge graph embedding models [22]. First, each object represented in the tensor is associated with initial random embeddings; these embeddings are then updated during the learning process such that the interactions of embeddings (*i.e.* the scoring function) yields high values for true combinations and lower values otherwise. Different models were developed to perform tensor decomposition where they vary in their modelling of the object embeddings, embedding interaction functions, and training objectives.

***Ranking training objectives.*** Tensor factorisation models traditionally use learning-to-rank loss functions as their training objectives. This allows them to perform tensor completion through ranking tensor combinations according to their factuality. They use different approaches for modelling ranking objectives such as pairwise and pointwise loss objectives. For example, the DistMult model uses a pointwise hinge loss function. Its objective is then to minimise the marginal difference between negative and positive scores [23], therefore, this makes the scores of positive combinations always higher than the scores of the negative combinations as shown in Eq. 1. On the other hand, other tensor factorisation models such as the ComplEx model [24] uses a pointwise logistic objective which aims to minimise the difference between combination scores and their assigned targets.

## Materials

In this study, we use a drug target interaction dataset extracted from the Comparative Toxicogenomics Database (CTD) [12, 13]. The CTD database contains data on chemicals, pathways, diseases, exposures, genes and phenotypes. It also contains different types of associations between these entities. In our study, we only consider the chemical gene associations where we filter out the interactions according to the related species to keep only the interactions assigned to humans.

We build a new benchmarking dataset, CTD38E, which contains associations between chemicals and their human protein targets from the CTD data. It also includes the different effect types related to these associations between the chemicals and proteins. The dataset includes a set of 38 different chemical effects which are filtered according to their coverage where we only keep effects that have 500 instances or more. We further divide the dataset into training and testing splits with $90\%$ and $10\%$ ratios respectively for the training and evaluation pipeline.

The different chemical effects describe an increase, decrease or uncategorised effect on the different protein attributes and activities such as methylation, oxidation, etc. For example, a chemical effect on a protein can increase the protein expression, decrease its abundance, have a general effect on its binding, etc.

The generated CTD38E dataset have a variable coverage of chemical effects where the change of protein expression associated effects have the highest coverage. Also, 32 out of the 38 represented chemical effects have approximately $\approx 15k$ or less instances in the dataset. On the other hand, the remaining effects have variable coverage that varies from $\approx 20k$ to $\approx 186k$ instances.

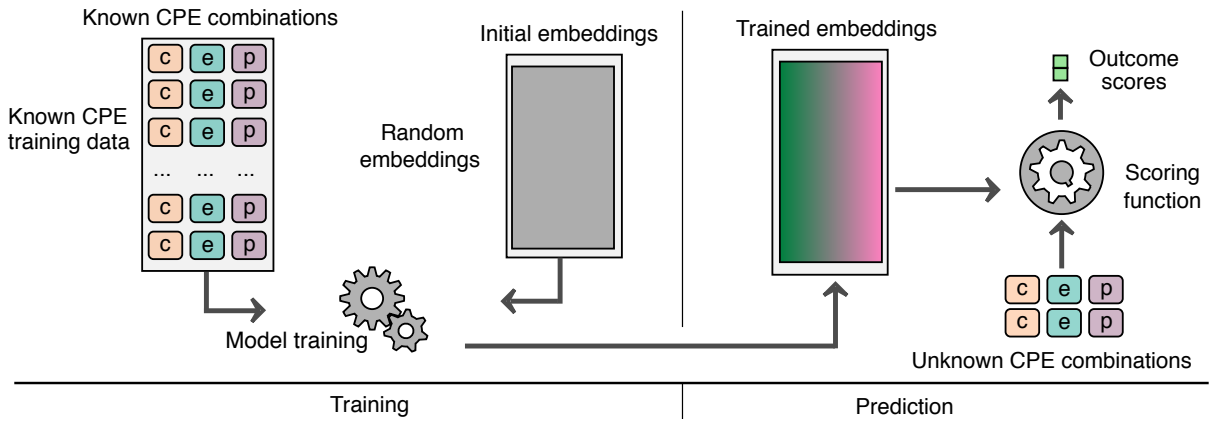


Figure 1: Summary of training and evaluation pipeline of the TriVec tensor factorisation approach. The abbreviation CPE denotes chemical protein effect.

## Methods

In this work, we use the TriVec tensor factorisation approach which provides an efficient means for modelling 3D tensors using multi-part embeddings. It also uses a multi-component embedding interaction function and multi-class training objective. In the following, we discuss the design of the TriVec tensor factorisation method and its training and evaluation pipelines. We further discuss its different unique properties such as its scoring function, training objective and embedding representation.

***The training and prediction pipelines.*** The tensor factorisation process operates by learning vector representations for the different tensor objects during the training phase. These representations are then used to predict the probability of unknown tensor combinations. (Figure 1) presents an illustration of the training and prediction pipelines of our approach, the TriModel.

In the training phase, our approach starts with consuming the set of known tensor combinations and an initial set of embeddings of the tensor objects as shown in (Figure 1). The method then iteratively processes the known tensor combinations to update the initial embeddings. The updates to the embeddings vectors are executed through a batch-based gradient decent optimisation procedure [(25)]. The objective of this optimisation procedure is to maximise the scores of the given true combinations and to minimise the scores of other random combinations as specified in (Eq. 1). This objective is dependent on the methods scoring function, *i.e.* embedding interaction function, where this function provides a score for each tensor combination using the embeddings of its objects.

The consumed known combinations, in this context, represent the training split of the CTD38E dataset where the method tries to learn an efficient representation for each chemical, effect and protein in the dataset. After a specific given number of training iterations, the method stores the current values of the tensor objects' embeddings. In the prediction phase, the method is able to provide a score for a given (*chemical*, *effect*, *protein*) combination using the learnt embeddings vectors. This is achieved by processing the embeddings corresponding to the combination's objects through the same scoring function used in the training phase. This procedure is shown in prediction part in (Figure 1).

***Multiple vector embeddings.*** The ComplEx model [(24)] has introduced the use of multiple vectors to represent single tensor objects. This allowed it to encoding both ordered and unordered tensor combinations. In our work, we use the TriModel embeddings [(26)] which is a similar approach that utilizes three embedding vectors for each tensor object. This enables efficiently encoding ordered combination like the ComplEx model with higher accuracy due to the extended representation.

***Modelling embedding interactions.*** The embedding interaction function, the scoring function, of the TriModel uses a combination of symmetric and asymmetric products to encode embedding interactions. This allow the method to

efficiently model both symmetric and asymmetric tensor combinations. The scoring function of the TriModel is defined as follows:

$$\eta^{\mathbf{M}}_{(c,e,p)} = \sum_{i=1}^{k} \overset{1}{\nu}_c \cdot \overset{1}{\nu}_e \cdot \overset{3}{\nu}_p + \overset{2}{\nu}_c \cdot \overset{2}{\nu}_e \cdot \overset{2}{\nu}_p + \overset{3}{\nu}_c \cdot \overset{3}{\nu}_e \cdot \overset{1}{\nu}_p, \tag{2}$$

where $\overset{1}{\nu}_c$, $\overset{2}{\nu}_c$ and $\overset{3}{\nu}_c$ are the three vector representations of the object $c$ and $k$ is the size of the single embedding vectors. This procedure can be simplified as a collection of three products $(\overset{1}{\nu}_c \cdot \overset{1}{\nu}_e \cdot \overset{3}{\nu}_p)$, $(\overset{2}{\nu}_c \cdot \overset{2}{\nu}_e \cdot \overset{2}{\nu}_p)$ and $(\overset{3}{\nu}_c \cdot \overset{3}{\nu}_e \cdot \overset{1}{\nu}_p)$. The first and third products, then, are asymmetric while the second product is a symmetric procedure.

***Training objective.*** The training objective of the TriVec model is to maximise the score of the true combinations and to minimise the scores of the other combinations as indicated in (Eq. 1). This is achieved through a multi-class loss objective that tries to maximise the score of each true combination compared to all possible corruption of its sides. For example, a true combination like (*27-hydroxycholesterol*, *increases the expression of*, *CCN5*) is compared to all combinations in the format of ($\cdots$, *increases the expression of*, *CCN5*) and (*27-hydroxycholesterol*, *increases the expression of*, $\cdots$) where the dots represents all the possible values of its category. This procedure is executed using the negative-log softmax loss that is defined as follows:

$$\begin{aligned} \mathcal{J}_{\texttt{TriModel-MC}} = & \sum_{c,e,p} [-2 \cdot \eta^{\mathbf{M}}_{(c,e,p)} + \log(\sum_{i'} \eta^{\mathbf{M}}_{(i',e,p)}) + \log(\sum_{k'} \eta^{\mathbf{M}}_{(c,e,p')}) \\ & + \frac{\lambda}{3} \sum_{m=1}^{M} \sum_{d=1}^{3} (|\overset{d}{\nu}_i| + |\overset{d}{\nu}_j| + |\overset{d}{\nu}_k|)], \end{aligned} \tag{3}$$

where $c'$ and $k'$ represent all possible chemicals and proteins respectively, $\lambda$ is a configurable weight parameter and the term $\frac{\lambda}{3} \sum_{m=1}^{M} \sum_{d=1}^{3} (|\overset{d}{\nu}_i| + |\overset{d}{\nu}_j| + |\overset{d}{\nu}_k|)$ is a regularisation term that represents the nuclear 3-norm [27] that is used for model generalisation purposes. This loss allows tensor factorisation methods to provide high accuracy predictions [27], however, it has limited scalability [26]. This occurs since the function processes the full vocabulary of tensor objects for each training instance. Therefore, it has a a quadratic space and time complexity unlike the traditional ranking objectives that have linear time and space complexity [23, 24, 28].

We also assess the performance of the TriVec method using a pointwise logistic loss function which enables highly scalable predictions compared to the previous multi-class loss approach. It is, however, known to have inferior accuracy when compared to the multi-class softmax loss. The pointwise logistic loss function is defined as follows:

$$\mathcal{J}_{\texttt{TriModel-Pt}} = \sum_{c,e,p} \log(1 + \exp[-l_{(c,e,p)} \cdot \eta^{\mathbf{M}}_{(c,e,p)}]), \tag{4}$$

where $l_{(c,e,p)}$ denotes the true label of the combination $(c, e, p)$ which is equal to 1 if the combination is true and 0 otherwise.

## Results

In this section, we discuss the design and configuration of our experiments. We first present the setup of our model training strategy: the models training configurations and the details parameters grid search. We then discuss our evaluation protocol, evaluation metrics and benchmarking data configuration. Finally, we compare the outcome evaluation results of the TriVec model to other approaches in terms of both the accuracy and efficiency.

***Experimental setup.*** In our experiments, we use the CTD38E benchmarking dataset which we have generated. The dataset is divided into two splits: training and testing. We divide the training split into two random splits for training and validation (90% for training and 10% for validation). The testing split is only used of the evaluation of investigated models.

| Model | MRR | Hits@10 | AUC-ROC | | | AUC-PR | | | Iter. time | Tot. time |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | N01 | N10 | N50 | N01 | N10 | N50 | (Sec.) | (Min.) |
| TransE | 0.07 | 0.20 | **0.99** | **0.99** | **0.99** | **0.99** | 0.93 | 0.82 | **1** | 19 |
| DistMult | 0.10 | 0.20 | 0.97 | 0.91 | 0.96 | 0.97 | 0.91 | 0.83 | **1** | **18** |
| ComplEx | 0.09 | 0.19 | 0.98 | 0.93 | 0.97 | 0.98 | 0.93 | 0.85 | 2 | 37 |
| TriVec - Pt | 0.09 | 0.20 | 0.98 | 0.98 | 0.98 | 0.99 | 0.94 | 0.86 | 19 | 33 |
| TriVec - MC | **0.28** | **0.41** | **0.99** | **0.99** | **0.99** | **0.99** | **0.95** | **0.88** | 32 | 151 |

Table 1: A Comparison between the TriVec model and other models model in terms of the mean reciprocal rank, Hits@10, area under the ROC and precision-recall curves, the training runtime for each training iteration and the total training runtime.

We compare the TriVec model to other tensor factorisation methods such as the DistMult and ComplEx models. We also compare it to the TransE model [28] which is a distance-based graph embedding model which can be utilised to perform tensor completion. We train all models through a grid search procedure to find the best hyperparameters for each model. The search space of the hyperparameters is defined as follows: the learning rate $\mathrm{lr} \in \{0.1, 0.3, 0.5\}$, embeddings size $k \in \{50, 75, 100, 150, 210\}$ and batch size $b \in \{1000, 3000, 5000, 8000\}$. The rest of the grid search hyper parameters are defined as follows: in the ranking loss approach, we use the negative sampling ratio $n \in \{2, 5, 10\}$, and in the multi-class approach we use the regularisation weight $\lambda \in \{0.1, 0.3, 0.35, 0.01, 0.03, 0.035\}$ and dropout $d \in \{0.0, 0.1, 0.2, 0.01, 0.02\}$. In the ranking loss approach, the number of training epochs is fixed to 1000. On the other hand, the number of epochs in the multi-class loss is 250.

***Evaluation protocol.*** We use the testing split of the CTD38E dataset to assess the predictive accuracy and required training runtime of each of the investigated methods. To assess the predictive accuracy, we use different types of metrics such as mean reciprocal rank and Hits@10, area under the ROC and precision recall curves. In the following, we give a short description of each of these metrics.

- *Mean reciprocal rank (MRR).* A ranking metric that is specified in assessing the quality of the highest predicted item in a rank. In our study, we use the MRR metric to assess if the model is able to find the right chemical and protein of a combination (in the testing split) in the set of all possible chemicals and drugs. This procedure resembles the link prediction procedure for knowledge graph completion [28].
- *Hits@10.* A ranking metric that has the same corruption mechanism and negative to positive ratio as the MRR metric. However, the Hits@10 focuses on the top 10 ranked items where it is equal to one if the item is found in the top 10 rank and zero otherwise. In our experiments, the reported Hits@10 values are the average of all the Hits@10 values of each testing combination.
- *The area under the ROC and precision recall curves (AUC-ROC and AUC-PR).* We use both these metrics with different positive to negative ratios to evaluate the models' sensitivity and specificity. We use three ratio, 1:1, 1:10 and 1:50 respectively, where the negative samples are generated randomly for each of the investigated chemical effects. The randomly generated negatives are filtered such that they do not intersect with any of the known combinations of the investigated effect. The AUC-ROC and AUC-PR metrics are then computed per chemical effect and averaged on the overall number of effects available in the data testing split.

***Implementation details.*** We use TensorFlow framework (GPU) on Python 3.5 to perform our experiments. All experiments were executed on a Linux machine with processor Intel(R) Core(TM) i70.4790K CPU @ 4.00GHz, 32 GB RAM, and an nVidia Titan Xp GPU. We have published the dataset and training logs; a set of model predictions are published in a figshare repository at: `https://figshare.com/articles/CTD-experiment/9383918`.

The source of our experiments is also published at: `https://github.com/samehkamaleldin/ecpi`.

***Comparison with other approaches.*** (Table 1) presents a comparison between the TriVec method and other approaches in terms of both predictive accuracy metrics and training runtime. The results show that the TriVec model achieves significantly better scores compared to other approaches in terms of the 1-versus-all negative to positive metrics (MRR and Hits@10). The results show that the TriVec model with the multi-class loss approach achieves the best results in terms of all the predictive accuracy metrics. For example, the results show that the TriVec -MC approach achieve 0.28 MRR score which is approximately 200% better than the scores of its pointwise loss version, ComplEx, DistMult, TransE models. The Hits@10 score of the TriVec -MC method is also approximately 100% better than all the other approaches.

The multi-class version of the TriVec model also achieves the best results in terms of the area under the ROC and precision recall curves. However, the difference between its scores and the scores of other models is small (ranges from 1% to 2%) in all the negative sampling configurations. The results also show that the achieved enhancements of the TriVec model positively correlate with the negative to positive ratio. This shows that the TriVec model is able to provide better results than other approaches in harder evaluation settings (N50), which can be supported by its results on the 1-versus-all evaluation metrics (MRR and Hits10).

Despite the predictive accuracy enhancements achieved by the multi-class loss version of the TriVec model, it requires significantly higher training time compared to all other approaches as shown in (Table 1). In this context, the results show that the TransE and DistMult model require the least training time compared to all other methods.

**Discussion**

In this section, we discuss in details the findings of our experiments and the details of evaluation scores of the highest performing methods in terms of the AUC-PR for each of the investigated chemical effects. We also discuss the different properties and features that the family of tensor factorisation methods enable. We then discuss the challenges and limitations associated with using tensor factorisation techniques for predicting the effects of chemicals on the human proteins. Finally, we discuss the intended future activities that we intend to perform in upcoming works to extend the scope and objective of this study.

***Scalability.*** Tensor factorisation methods are representation learning techniques that operate by learning efficient vector representations of tensor objects. They then use representations to assess the factuality of tensor combinations. The embedding learning procedure is known to operate with linear time and space complexity [24, 26]. This allows tensor factorisation methods to provide scalable predictions compared to other traditional approaches that require more complex feature processing routines.

Furthermore, the predictive procedure of the tensor factorisation methods is a constant time complexity routine ($\mathcal{O}1$). This gives them a significant scalability advantage over other approaches that require feature processing in their predictive procedure after training.

***Significance to clinical research.*** Despite the high accuracy of computational approaches in multiple biological inference tasks, they are never supposed to replace clinical experimentation. They however, aim to assist researchers in biological studies in prioritizing their experimentation configurations. For example, our study aims to assist biologists who are experimenting on different chemical substances to assess their effects on human proteins. Our proposed computational approach provides predictions that enables ranking possible configurations (combinations) of chemicals, proteins and their associated chemical effects according to their likelihood of being present. Biologists can use this rank to prioritize the order of executed experimentation to focus on the highest ranked combinations.

***Limitations.*** Despite the high predictive accuracy of the tensor factorisation approaches, they are not easily interpreted. These methods operate as black boxes where it is hard to determine which set of features have affected their predictions. This also affects the trust in their predictions, especially in the biomedical domain, as it is critical to understand the rationale behind predictions to avoid misinformed judgements.

Tensor factorisation procedures build representations of tensor objects using their existing known combinations. Therefore, they provide low quality representations of the under represented objects [22]. In the context of biological information, the coverage of biological entities has a high variance due to the unbalanced focus of clinical and research studies of biological entities such as proteins, drugs, etc. Therefore, this affects the quality of representations of tensor factorisation methods of the under represented objects. In addition, the tensor factorisation methods are unable to provide beneficial representation of newly introduced objects *e.g.* new chemicals and proteins, as they require prior information to operate.

***Future works.*** In future works, we intend to incorporate the information about the associated body tissues of each chemical–protein interactions. This will enable more accurate and specialised effects since the interactions between chemicals and proteins are strongly affected by the associated tissue context.

We also aim to experiment with representation learning methods that utilize protein and chemical structures rather than their prior information. This direction will enable more accurate and efficient predictions for new and under studied chemicals and proteins.

## Conclusions

In this work, we have studied the problem of identifying the effects of the interactions between chemicals and human proteins. We have shown the importance of computational methods in assisting clinical research in this particular task. We have then proposed using tensor factorisation methods to predict the effects of chemicals on human proteins where we modelled the chemicals, their effects and the proteins as a tensor. We then used tensor factorisation to learn efficient representations of the tensor objects to be able to predict new combinations.

We have adopted the TriVec method as our main approach, and we have built a benchmarking dataset (CTD38E) based on the comparative Toxicogenomics database to train and evaluate our investigated approaches. We have then shown by computational experimental evaluation that the TriVec method outperforms other known tensor factorisation methods in the studied task in terms of different evaluation metrics such as the MRR, Hits@10 and the area under the ROC and precision recall curves.

Finally, we have discussed different properties and limitations of the different tensor factorisation approaches, and we have presented the set of intended future activities to extend the scope and objective of this study.

## Funding

The work presented in this paper was supported by the CLARIFY project funded by European Commission under the grant number 875160, and by the Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland (supported by the Science Foundation Ireland grant (12/RC/2289_P2).

## Refrences

1. Marnie L. Macdonald, Jane E Lamerdin, Stephen F. Owens, Brigitte H. Keon, Graham K. Bilter, Zhidi Shang, Zhengping Huang, Helen Yu, Jennifer M. Dias, Tomoe Minami, Stephen W. Michnick, and John K. Westwick. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nature Chemical Biology*, 2:329–337, 2006.

2. Dirk Brehmer, Zoltán Greff, Klaus Godl, Stephanie Blencke, Alexander Kurtenbach, Martina Weber, Stephan Müller, Bert Klebl, Matthew Cotten, Gy. Kéri, Josef Wissing, and Henrik Daub. Cellular targets of gefitinib. *Cancer research*, 65 2:379–82, 2005.

3. Brent M Kuenzi, Lily L. Remsing Rix, Paul Alexander Stewart, Bin Fang, Fumi Kinose, Annamarie T Bryant, Theresa A Boyle, John Matthew Koomen, Eric B. Haura, and Uwe Rix. Polypharmacology-based ceritinib repurposing using integrated functional proteomics. *Nature chemical biology*, 13 12:1222–1231, 2017.

4. Kristiina Iljin, Kirsi Ketola, Paula Vainio, Pasi K Halonen, Pekka Kohonen, Vidal rer. nat. Fey, Roland C Grafström, Merja Perälä, and Olli Kallioniemi. High-throughput cell-based screening of 4910 known drugs and drug-like

small molecules identifies disulfiram as an inhibitor of prostate cancer cell growth. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15 19:6070–8, 2009.

5. Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. In *Bioinformatics*, 2009.

6. Rawan S. Olayan, Haitham Ashoor, and Vladimir B. Bajic. Ddr: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. In *Bioinformatics*, 2018.

7. Sameh K Mohamed, Vít Nováček, and Aayah Nounu. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 08 2019.

8. Sameh K. Mohamed, Vít Novácek, and Pierre-Yves Vandenbussche. Knowledge base completion using distinct subgraph paths. In *SAC*, pages 1992–1999. ACM, 2018.

9. Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. In *Bioinformatics*, 2018.

10. Emir Muñoz, Vít Novácek, and Pierre-Yves Vandenbussche. Using drug similarities for discovery of possible adverse reactions. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*. AMIA, 2016.

11. Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. In *Bioinformatics*, 2017.

12. Carolyn J. Mattingly, Michael C. Rosenstein, Glenn T. Colby, John N. Forrest, and James Boyer. The comparative toxicogenomics database (ctd): a resource for comparative toxicological studies. *Journal of experimental zoology. Part A, Comparative experimental biology*, 305 9:689–92, 2006.

13. Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research*, 47(D1):D948–D954, 09 2018.

14. Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673, 2004.

15. Curtis Robert Chong and David J Sullivan. New uses for old drugs. *Nature*, 448:645–646, 2007.

16. Tudor I. Oprea, Alexander Tropsha, Jean-Loup Faulon, and Mark D. Rintoul. Systems chemical biology. *Nature chemical biology*, 3 8:447–50, 2007.

17. Mónica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321 5886:263–6, 2008.

18. Francesco Di Iorio, Timothy Rittman, Hong Ge, Michael P. Menden, and Julio Sáez-Rodríguez. Transcriptional data: a new gateway to drug repositioning? In *Drug discovery today*, 2013.

19. Philippe Sanseau, Pankaj Agarwal, M. R. Barnes, Tomi Pastinen, Jeremy B Richards, Lon R. Cardon, and Vincent E. Mooser. Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30:317–320, 2012.

20. Guangxu Jin and Stephen T. C. Wong. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug discovery today*, 19 5:637–44, 2014.

21. Ned Calonge, Diana B. Petitti, Thomas G Dewitt, Allen J. Dietrich, Kimberly D. Gregory, Richard J. Harris, George T. Isham, Michael L Lefevre, Roseanne M. Leipzig, Carol J. Loveland-Cherry, Lucy Nelle Marion, Bernadette Mazurek Melnyk, Virginia A. Moyer, Judith K. Ockene, George F. Sawaya, and Barbara P. Yawn. Screening for colorectal cancer: U.s. preventive services task force recommendation statement. *Annals of internal medicine*, 149 9:627–37, 2008.

22. Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

23. Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, 2015.

24. Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

25. Phuong Thi Tran et al. On the convergence proof of amsgrad and a new version. *IEEE Access*, 7:61706–61716, 2019.

26. Sameh K. Mohamed and Vít Novácek. Link prediction using multi part embeddings. In *ESWC*, volume 11503 of *Lecture Notes in Computer Science*, pages 240–254. Springer, 2019.

27. Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 2869–2878. JMLR.org, 2018.

28. Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.